IPUMS Data Training Exercise:

IPUMS International
Data Extract and Analysis
(Exercise 2 for R)

IPUMS INTERNATIONAL

Learning Goals

- Understand how IPUMS International dataset is structured
- Create and download an IPUMS data extract
- Decompress the data file and read the data into a statistical package
- Analyze the demographic and population characteristics of Cambodia, Ireland, Uruguay using sample code
- Validate data analysis work using the answer key
- Understand how IPUMS data can be leveraged to explore research interests

Exercise Research Question and Variables

In this exercise, you will gain basic familiarity with the IPUMS International data exploration and extract system to answer the following research question: "What are the differences in in water supply, internet access, car ownership, and age distribution among Cambodia, Uruguay, and Ireland?" You will create a data extract that includes the variables WATSUP, SEX, INTRNET, AUTOS, EDATTAIN, AGE, HHWT; then you will use the sample code to analyze these data.

Register as an IPUMS International User

Go to http://international.ipums.org, click on User Registration and Login and Apply for access. On the login screen, enter email address and password and submit your application. Please note that IPUMS International user applications are reviewed by IPUMS staff, and a final decision may take 2-5 business days.

R Code to Review

This tutorial's sample code and answers use the so-called "tidyverse" style, but R has the blessing (and curse) that there are many different ways to do almost everything. If you prefer another programming style, please feel free to use it. However, for your reference, these are some quick explanations for commands that this tutorial will use:

Code	Purpose
%>%	The pipe operator which helps make code with nested function calls is easier to read. When reading code, read as "and then". The pipe make it so that the code like ingredient %>% stir() %>% cook() is equivalent to cook (stir(ingredients)) (read as "take *ingredients* and then *stir* and then *cook*").
as_factor	Converts the value labels provide for IPUMS data into a factor variable for R
summarize	Summarize a datasets observations to one or more groups
group_by	Set the groups for the summarize function to group by
filter	Filter the dataset so that it only contains these variables
mutate	Add on a new variable to a dataset
ggplot	Make graphs using ggplot2

weighted.mean	Get the weighted mean of the “a” variable

Make a Data Extract

- Navigate to the IPUMS International homepage and click on "Browse Data."

Select samples

- Click on the "Select Samples" button to choose the census samples to include in your extract.
- Check the boxes for the 2008 sample for Cambodia, 2006 for Ireland, and 2006 for Uruguay.
- Submit your sample selections by clicking the Submit sample selections box.
- Note that by selecting samples first, you will now only see variables available for Cambodia, Ireland, or Uruguay.
    - If you would prefer to see all variables, regardless of their availability in your selected samples, click on "Display Options" from the main variable browsing page, and choose to display variables that are not available in your selected samples.

Select variables

- The variable drop-down menus allow you to explore variables by topic. For example, you might find variables about occupational participation under the "Work" group.
- The search tool allows you to search for variables. Observe the options for limiting your search results by variable characteristics or variable type.
- Add a variable to your cart by clicking on the plus sign in the "Add to Cart" column of the topical variable list, or list of search results.
- View more information about the variable by clicking on the variable name, and navigating through the tabs that include a description of the variable, codes and value labels, the universe of persons asked the question, and information on the

comparability of the variable among other pieces of information. If you are reviewing variable-specific information, you may click on the "Add to Cart" button near the top of the screen to add this variable to your data cart.

- Use either the drop down menu or the search feature to select the following variables, and add them to your data cart.
    - WATSUP: Water supply
    - SEX: sex
    - INTRNET: Internet access
    - AUTOS: Automobiles available
    - EDATTAIN: Educational attainment
    - AGE: Age
    - HHWT: Household weight technical variable

Review data and request the extract

- Click on the "View Cart" button underneath your data cart.
- Review your variable and sample selection to ensure your extract will be complete.
    - You may notice a number of additional variables you did not select are in your cart; IPUMS preselects a number of key technical variables, which are automatically included in your data extract.
- Add additional variables or samples if they are missing from your extract, or click the "Create Data Extract" button.
- Review the Extract Request screen that summarizes your extract; add a description of your extract (e.g., "Differences water supply, internet access, car ownership, and age distribution among Cambodia (2008), Ireland (2006), and Uruguay (2006)?") and click "Submit Extract".
- You will receive an email when your data extract is available to download.

Getting the Data Into Your Statistics Software

The IPUMS International extract builder provides raw ASCII data files and the command files necessary for reading the raw data into a stats package. Note that these instructions are for R. If you would like instructions for a different stats package, see https://www.ipums.org/exercises.shtml.

Download the data

- Follow the link in the email notifying you that your extract is ready, or by clicking on the "Download and Revise Extracts" link on the left-hand side of the IPUMS International homepage.
- Right-click on the data link next to the extract you created.
- Choose "Save Target As..." (or "Save Link As...")
- Save into your preferred working directory. This tutorial assumes you will save the file into "Documents" (which should pop up as the default location).
- Do the same thing for the DDI link next to the extract
- (Optional) Do the same thing for the R script
- You do not need to decompress the data to use it in R

Install the IPUMSR package

- Open R from the Start menu
- If you haven't already installed the ipumsr package, in the command prompt, type the following command: `install.packages ("ipumsr")`

Read in the data

- Set your working directory to where you saved the data above by adapting the following command (Rstudio users can also use the "Project" feature to set the working directory. In the menubar, select File->New Project->Existing Directory and then navigate to the folder):

`setwd("~/") # "~/"` goes to your Document director on most computers

- Run the following command from the console, adapting it so it refers to the extract you just created (note the number may not be the same depending on how many extracts you've already made):

```
library(ipumsr)
ddi <- read_ipums_ddi("ipums_00001.xml")
data <- read_ipums_micro(ddi)
```

#Or, if you downloaded the R script, the following is equivalent:
`# source("ipumsi_00001.R")`

- This tutorial will also rely on the dplyr and ggplot2 packages, so if you want to run the same code, run the following command (but if you know other ways better, feel free to use them):

```
library(dplyr)
library(ggplot2)
```

- To stay consistent with the exercises for other statistical packages, this exercise does not spend much time on the helpers to allow for translation of the way IPUMS uses labelled values to the way base R does. You can learn more about these in the value-labels vignette in the R package. From R run command:

```
vignette("value-labels", package + "ipumsr")
```

Analyze the Data

Part 1: Variable documentation

For each variable below, search through the tabbed sections of the variable description to answer each question.

1. Find the codes page for the SAMPLE variable and write down the code values for:

a. Cambodia 2008 _____

b. Ireland 2006 _____

c. Uruguay 2006 _____

2. Are there any differences in the universe of WATSUP among the three samples?

_____

3. What is the universe for EMPSTAT in:

a. Cambodia 2008?_____

b. Ireland 2006? _____

c. Uruguay 2006? _____

Part 2: Frequencies

4. How many individuals are in each of the sample extracts?

a. Cambodia 2008 _____

b. Ireland 2006 _____

c. Uruguay 2006 _____

```
data %>%

  group_by(SAMPLE = as_factor(SAMPLE, level = "both")) %>%
  summarize(n = n())
```

To get a more accurate estimation of demographic patterns within a country from the sample, you will have to turn on the person weight.

5. Using weights, what is the total population of each country?

a. Cambodia 2008 _____

b. Ireland 2006 _____

c. Uruguay 2006 _____

```
data %>

 group_by(SAMPLE = as_factor(SAMPLE)) %>%
 summarize(n = sum(PERWT))
```

6. Using weights, what proportion of individuals in each country did not have access to piped water?
    a. Cambodia 2008 _____
    b. Ireland 2006 _____
    c. Uruguay 2006 _____

```
data %>%
   mutate (
     NOT_PIPED = WATSUP %>%
        lbl_collapse(~.val %/% 10) %>%
        as_factor() %>%
        {.! = "Yes, piped water"}
   ) %>%
   group_by(SAMPLE = as_factor(SAMPLE)) %>%
   summarize (NOT_PIPED = weighted.mean(NOT_PIPED, PERWT, na.rm
= TRUE
```

Part 3: Weighted frequencies (HHWT)

Suppose you were interested in the number of people with or without water supply, but in the number of households. You will need to use the household weight.

In order to use household weight, you should be careful to select only one person from each household to represent that household's characteristics. You will need to apply the household weight (HHWT) to identify only one person from each household. Use the filter statement to select only cases where the PERNUM equals 1.

7. What proportion of households in each country did not have access to piped water?
    a. Cambodia 2008 _____
    b. Ireland 2006 _____

c. Uruguay 2006 _____

```
data %>%
  filter(PERNUM == 1) %>%
  mutate (
    NOT_PIPED = WATSUP %>%
      lbl_collapse(~.val %>% 10) %>%
      as_factor() %>%
      {.! = "Yes, piped water"}
  )%>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(NOT_PIPED = weighted.mean(NOT_PIPED, HHWT, na.rm =
TRUE))
```

8. In what country do individuals have the most access to the internet?

   _____

```
data %>%
  mutate (
    HAVE_INTERNET = INTERNET %>%
      as_factor() %>%
      {. == "Yes"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(HAVE_INTERNET + weighted.mean(HAVE_INTERNET, PERWT,
na.rm = TRUE))
```

9. In that country, what proportion of households have both access to internet and at least one car? _____

Part 4: Trends

Note: First, you'll have to generate a dummy variable that is "1" when the household has at least one car and internet and "zero" in all other cases.

```
data %>%
  filter(as_factor(SAMPLE == "Ireland 2006") %>%
  mutate(
    HAVE_INTERNET = INTERNET %>%
      as_factor() %>%
      {. == "Yes"},
    HAVE_AUTO = AUTOS %>%
    {. > 0 & . <8}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(HAVE_BOTH = weighted.mean(HAVE_INTERNET & HAVE_AUTO,
PERWT, na.rm = TRUE))
```

10. In which country is educational attainment (Secondary and University in particular) between men and women most equal? Least equal?
    a. Most equal completion rates: _____
    b. Lease equal completion rates: _____

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(
    HAVE_SEC = weighted.mean(EDATTAIN == 3, PERWT, na.rm =
TRUE),
    HAVE_UNIVE = weighted.mean(EDATTAIN == 4, PERWT, na.rm =
TRUE)
)
```

Part 5: Graphical Analysis

11. Approximately what percent of Uruguay's population is around 50 years old?

    _____

12. Compare the age distributions of Cambodia and Ireland. Is this a pattern that could be observed in other developed and developing nations? _____

13. Can the shape of the histogram of Ireland compared to the other countries indicate anything about the differences in data collection?

```
ggplot(data, aes(x = as.numeric(AGE), y = ..prop..,weight =
PERWT)) +
  geom_bar() +
  facet_wrap(~as_factor(SAMPLE), ncol = 1)
```

14. What (approximately) are the median ages for men and women in each of these countries?

    a. Women:

       Cambodia 2008 _____ Ireland 2006 _____ Uruguay 2006 _____

    b. Men:

       Cambodia 2008 _____ Ireland 2006 _____ Uruguay 2006 _____

```
Data_summary <- data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(age_med = median(AGE))

Ggplot(data_summary, aes(x = SAMPLE, y = age_med, fill = SEX)) +
  geom_col(position = "dodge", width = 0.8) +
  scale_fill_manual(values = c(Male = "#7570b3", Female =
"#e6ab02"))
```

Answers

Part 1: Variable Documentation

1. Find the codes page for the SAMPLE variable and write down the code values for:
    a. Cambodia 2008: 116200801
    b. Ireland 2006: 37220601
    c. Uruguay 2006: 858200621
2. Are there any differences in the universe of WATSUP among the three samples? Cambodia 2008: Regular households, Ireland 2006: Private households in non-temporary dwellings, Uruguay 2006: All households. All have technical differences, Uruguay being most inclusive, and Ireland being the most precise.
3. What is the universe for EMPSTAT:
    a. in Cambodia 2008? All persons.
    b. in Ireland 2006? Non-absent persons age +15.
    c. Uruguay 2006? Persons age 14+.

Part 2: Frequencies

4. How many individuals are in each of the sample extracts?
    a. Cambodia 2008: 1,340,121
    b. Ireland 2006: 440,314
    c. Uruguay 2006: 256,866
5. Using weights, what is the total population of each country?
    a. Cambodia 2008: 13401210
    b. Ireland 2006: 4403140

c. Uruguay 2006: 3065604

6. Using weights, what proportion of individuals in each country did not have access to piped water? *(Note that we have treated NIU/Unknown as lacking water, but it would also be reasonable to treat them as missing.)*

   a. Cambodia 2008: 85.68%

   b. Ireland 2006: 5.61%

   c. Uruguay 2006: 3.22%

Part 3: Weighted Frequencies

7. What proportion of households in each country did not have access to piped water? *(Note that we have treated NIU/Unknown as lacking water, but it would also be reasonable to treat them as missing.)*

   a. Cambodia 2008: 86.51%

   b. Ireland 2006: 9.90%

   c. Uruguay 2006: 3.28%

8. In which country do individuals have the most access to the internet?
   Ireland (*53.1%, including NIU/Unknown as not having access.*)
9. In that country, what proportion of households have both access to internet and at least one car? 50.6% (*including NIU/Unknown as not having access*)

Part 4: Trends

10. In which country is educational attainment (Secondary and University in particular) between men and women most equal? Least equal?

    a. Most equal completion rates: Uruguay (18.7%/19.8%; 4.0%/4.2%)

b. Least equal completion rates: Cambodia (4.7%/2.4%; 1.3%/0.6%) *(again including NIU/Unknown as not that level of education)*

Part 5: Graphical Analysis

11. Approximately what percent of Uruguay's population is around 50 years old? ~2.4%
12. Compare the age distributions of Cambodia and Ireland. Is this a pattern that could be observed in other developed and developing nations? A large proportion of Cambodia's population is 25 or younger, while the mean age of Ireland's population seems a bit older.
13. Can the shape of the histogram of Ireland compared to the other countries indicate anything about the differences in data collection? All Ireland samples provide single years of age through 19 and 5-year age intervals thereafter, top-coded at 85+ (from the Comparability Tab on the website)
14. What (approximately) are the median ages for men and women in each of these countries?

    a. Women:

    Cambodia 2008: 23 Ireland 2006: 32 Uruguay 2006: 35

    b. Men:

    Cambodia 2008: 20 Ireland 2006: 32 Uruguay 2006: 32